

# **GOlorize User Guide**

## **for Cytoscape version 2.4.x and BiNGO version 2.x**

GOlorize is a Cytoscape plug-in for advanced network visualization, which uses Gene Ontology (GO) categories as a source of class information to direct the layout process and to emphasize the biological function of the nodes. The implementation of GOlorize version 2.4 is based on the BiNGO plug-in, version 2.x, an efficient tool to determine the GO categories that are overrepresented in a selected part of a given network. The both plug-ins are used within Cytoscape, which is an open source bioinformatics software platform for visualizing and integrating molecular interaction networks. The main advantage of GOlorize compared to other graph layout tools is the possibility to incorporate the GO class information already in the node placement phase using a modified version of the force-directed layout algorithm. An extra attraction force is mediated through additional class nodes which represent the GO categories of interest. This version of GOlorize enables also the usage of other node attribute information than GO categories when defining the node classes of interest. GOlorize takes full advantage of Cytoscape's sophisticated filtering, analysis and visualization properties, allowing the user to produce customized high-quality network images.

## **Installation**

This version of GOlorize and the user-guide is compatible only with the Cytoscape, version 2.4., which can be freely downloaded from the Cytoscape project website <http://www.cytoscape.org/>. If not already installed on the computer, download and install also the Java 2 Runtime Environment, version 1.5 or higher, from <http://www.java.com/en/download/index.jsp>. Place the GOlorize2-4.jar file into the local Cytoscape /plugins directory. Note that in this version the BiNGO annotation files are incorporated directly in the GOlorize.jar file, and the version does not support the import ontology and annotation mechanism of Cytoscape 2.4 in the BiNGO over-representation analysis. Therefore, for updating the NCBI-based GO annotations, open the GOlorize.jar file with your favorite archive manager, and replace the original annotation files with the new files copied from the updated BiNGO.jar file. The new BiNGO.jar files can be downloaded from <http://www.psb.ugent.be/cbd/papers/BiNGO/index.htm>. Alternatively, the GO Consortium annotation files can be updated by substituting the old annotation or ontology files with their new versions downloaded directly from the Gene Ontology project homepage <http://www.geneontology.org>.

## **Tutorial**

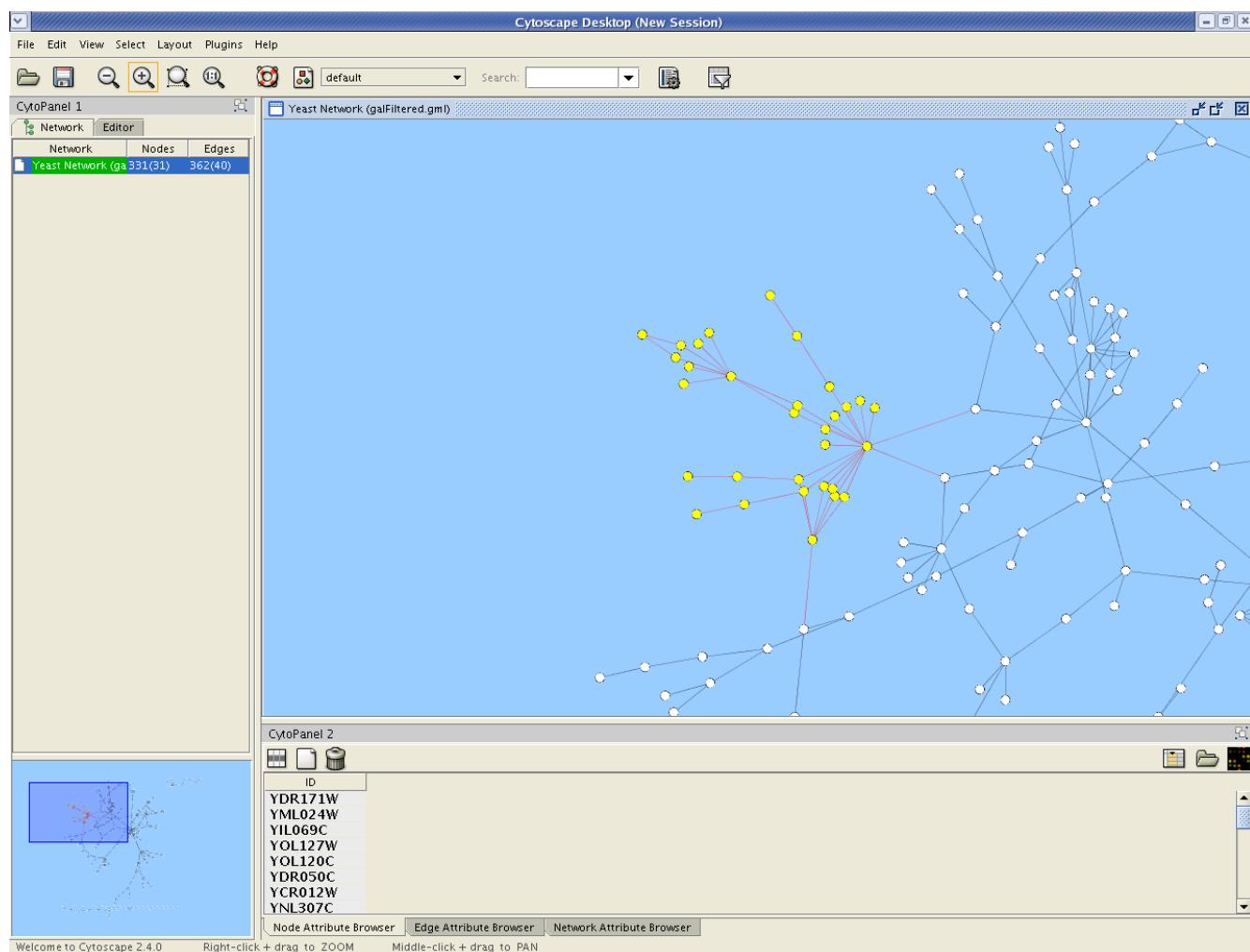
The usage of the plug-in is demonstrated with the following walk-through example.

## Step 1 - Starting the plug-in

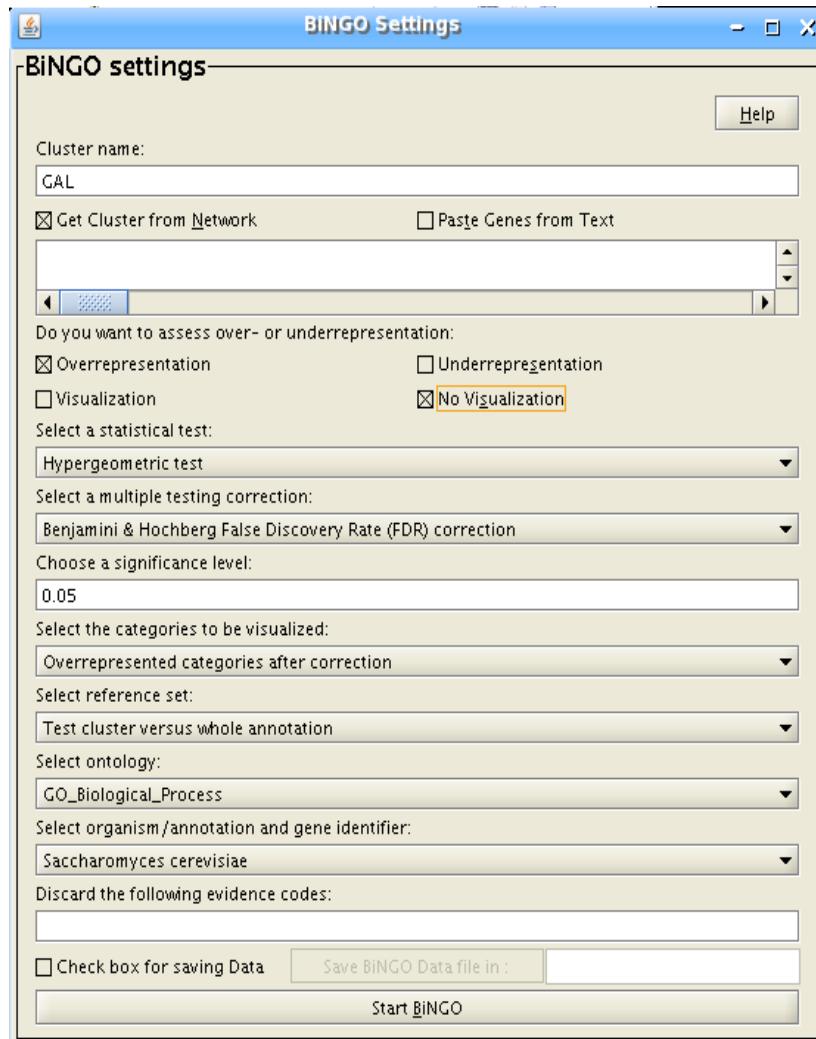
Start Cytoscape, version 2.4.x, and use the File menu from the Cytoscape main window to import the example network galFiltered.xgmml from the sampleData folder. Select GOlorize from the Cytoscape Plugins menu to start the interactive layout process. There are two modes how to define the node classes being used in the network visualization. The default mode corresponds to finding overrepresented GO categories in the selected network using BiNGO, version 2.x, whereas the alternative mode allows the user to define the node classes based on other attributes, such as expression data, or attributes imported from external files. The default mode is demonstrated in detail in Steps 2-5, and some illustrative examples of the attribute-based class definitions are given in Step 6.

## Step 2 – Using the BiNGO settings

Select a cluster of nodes in the network view, indicated in yellow nodes, which will be used as an input in the determination of GO categories that are statistically overrepresented in the network (see an example screen-shot below). Click on the Start BiNGO button in the GOlorize window.



The BiNGO Settings panel pops up. In this panel, the user can specify several parameters for the GO over-representation analysis, such as the type of a statistical test and multiple testing correction used, as well as the significance level, e.g.  $p < 0.05$ , which controls the number of enriched categories that will be outputted. For more information about the BiNGO settings and its operation, please see the BiNGO manual at <http://www.psb.ugent.be/cbd/papers/BiNGO/manual.htm>. The particular selections for our example case are shown below. Press Start BiNGO button to proceed the example.



### Step 3 – Choosing the GO categories

Having parsed the annotations and calculated the tests and their corrections, the BiNGO results appear in the GOLORIZE tab named after the BiNGO Settings Cluster name (GAL in the example below). This table lists the GO categories overrepresented in the selected subnetwork. The columns include the GO-ID term of the category and its description, along with the original and corrected statistical significances (p-val and corr p-value), the number of nodes in the selected subnetwork and in the complete annotation that belong to the particular category (cluster freq and total freq), and the node IDs annotated to the category (either directly or to its parent categories).

GOlorize

Start BiNGO Attributes Apply coloring All nodes in view Coloring effect Default pie size Level of Details

Selected Layout GAL

GO,default,Saccharomyces cerevisiae/process Close

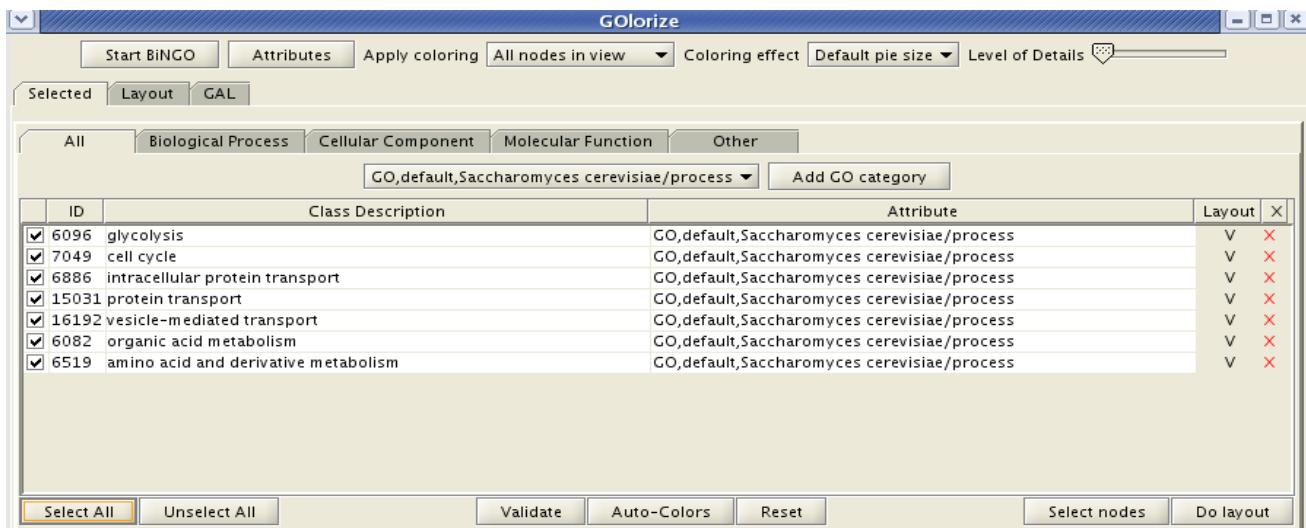
ID	Class Description	p-val	corr p...	cluster fr...	total freq	genes annotated to the term
6096	glycolysis	1.6 E-10	3.8 E-8	6/23 26...	22/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
6007	glucose catabolism	1.9 E-9	2.2 E-7	6/23 26...	32/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
19320	hexose catabolism	3.4 E-9	2.7 E-7	6/23 26...	35/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
46365	monosaccharide catabolism	8.0 E-9	4.6 E-7	6/23 26...	40/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
46164	alcohol catabolism	1.3 E-8	5.9 E-7	6/23 26...	43/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
16052	carbohydrate catabolism	7.2 E-8	2.4 E-6	6/23 26...	57/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
44275	cellular carbohydrate catabolism	7.2 E-8	2.4 E-6	6/23 26...	57/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
6006	glucose metabolism	1.5 E-7	4.3 E-6	6/23 26...	64/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
6092	main pathways of carbohydrate metabolism	3.2 E-7	8.4 E-6	6/23 26...	73/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
19318	hexose metabolism	8.1 E-7	1.9 E-5	6/23 26...	85/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
5996	monosaccharide metabolism	1.3 E-6	2.7 E-5	6/23 26...	92/5638...	YCR012W YCR254W YAL038W YDR050C YPL075W Y...
42257	ribosomal subunit assembly	1.8 E-6	3.5 E-5	5/23 21...	53/5638...	YPR102C YGR085C YLR075W YOL127W YML024W
6066	alcohol metabolism	2.4 E-6	4.2 E-5	7/23 30...	162/563...	YCR012W YOL086C YGR254W YAL038W YDR050C Y...
42255	ribosome assembly	3.9 E-6	6.6 E-5	5/23 21...	62/5638...	YPR102C YGR085C YLR075W YOL127W YML024W
						YCR012W YOL086C YGR254W YAL038W YDR050C Y...

Select All Unselect All Validate Reset Select nodes Do layout

The user can choose the categories that will be applied in the layout by checking the corresponding rows. It is also possible to select categories from multiple overrepresentation analyses, by re-starting the BiNGO again from the GOlorize panel, perhaps with different ontologies or parameter settings. Each BiNGO run is identified by its name in the tab list. The ontology being used is displayed above the categories (GO Biological Process of *Saccharomyces cerevisiae* ontology in the example case). After checking the categories of interest, press the Validate button and click the Selected tab.

In the Selected tab, all the GO categories selected from the BiNGO result(s) are shown. In this panel, the user can also manually add arbitrary GO categories by pressing the Add GO category button and typing the corresponding ID terms. In our example case, we have selected rather arbitrarily seven GO categories (shown below), which will be used in the network visualization process. Although class overlapping is allowed, we recommend not use more than 6 terms per node. It may be beneficial to select the categories of interest from the the higher-level terms, rather than using very broad categories.

Clicking on the GO Term ID number (the first column below) in the Selected tab opens an AmiGO web page for the particular GO category. The page contains a detailed view of information on the GO annotation and it also allows the user to browse, query and visualize the cross-links between the selected category and other available data from GO. The icons below the Layout column (V) indicate whether or not the corresponding categories will be used in the layout process, and the last column (X) can be used to exclude the categories from the subsequent node coloring and placement steps.



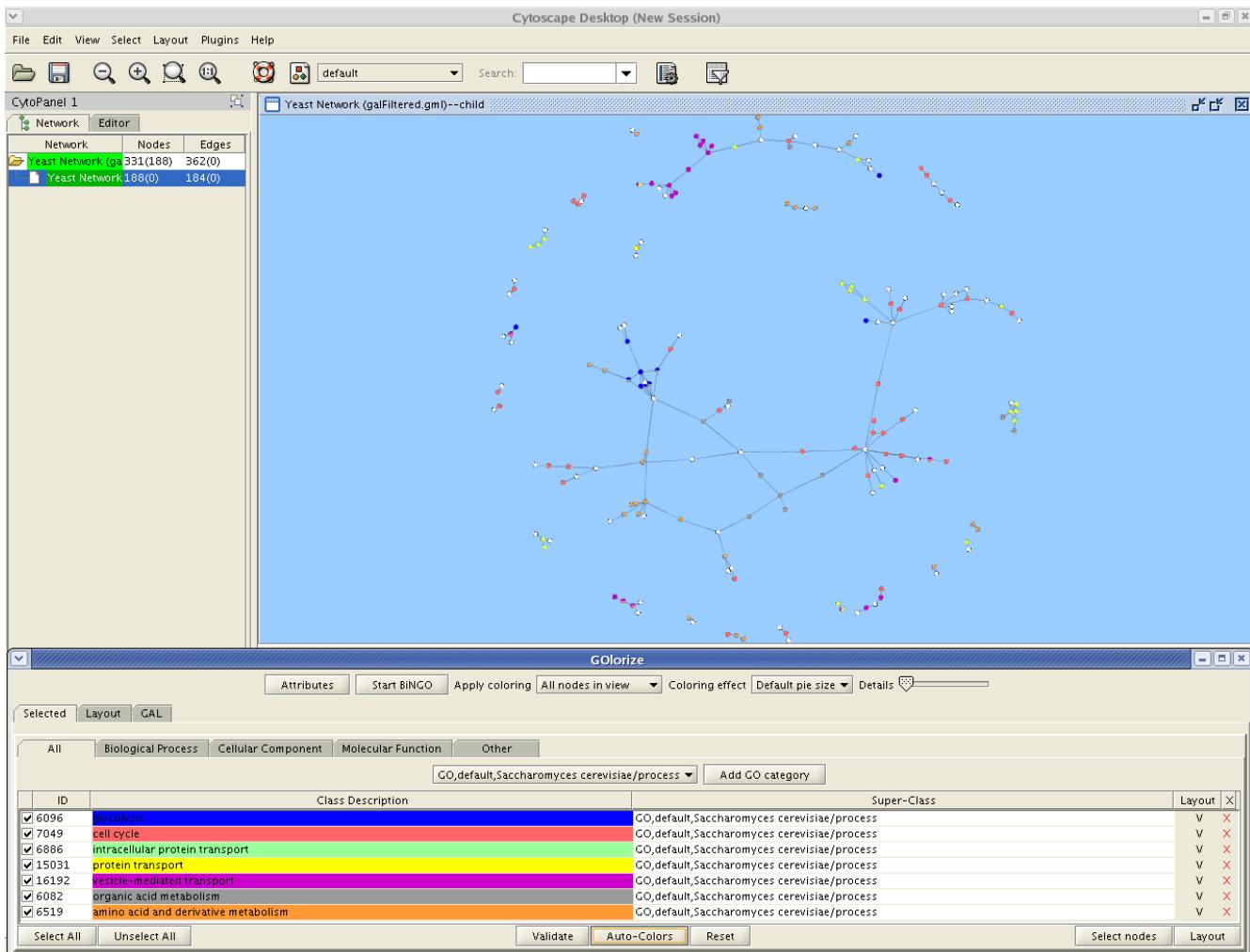
#### Step 4 – Coloring the selected nodes

If desired, the visualization can be focused only on the neighborhood of the selected categories by node selection mechanisms. The Select nodes button applies to the categories with the left-most check-boxes selected. It can be used in conjunction with other selection options in Cytoscape (choose Small pie size in the Coloring effect to see the selections in the main network view). In the example, we have selected all the nodes from the galFiltered network that belong to the seven categories, as well as their first neighbors, and added them into a new network (a subnetwork with 184 nodes and 181 edges).

Pressing the Auto-Colors button generates automatically a color-coding for the selected node classes. Alternatively, the user can manually choose the color of choice for each GO category. Due to hierarchical organization of the categories, each node can belong to none, unique or several classes. The unclassified nodes have the default node color of Cytoscape visualization, adjustable in Set Visual Style menu item. If a node belongs to several classes, a convenient pie coloring is applied. The user can also specify whether coloring is applied to all nodes in the network view or to selected nodes only.

Displaying and coloring the nodes with Golorize depends on two Cytoscape rendering properties, `render.nodeBorderThreshold` and `render.coarseDetailThreshold`, which can be manually modified from the `Edit->Preferences->Properties` menu. The Level of details slide bar provides an interactive way to adjust these parameters for the given network and for the selected zoom level. The maximum value (right) corresponds to those values needed to visualize all the colored pie charts in the whole network display. The minimum value (left) corresponds to the given default value of these parameters.

After validating the selections and applying the color-coding to the Cytoscape's Spring Embedded layout, the selected subnetwork view (galFiltered--child) should look something like the layout below (due to randomness in the layout process, the results from different runs of the Spring Embedded layouts may not be identical).



The above layout shows that while the color-coding can efficiently facilitate the visual interpretation of the class information, it is in many cases not enough for discovering whether or not the network contains a GO class structure superimposed on the underlying connection structure.

### Step 5 – Lay outing with GO classes

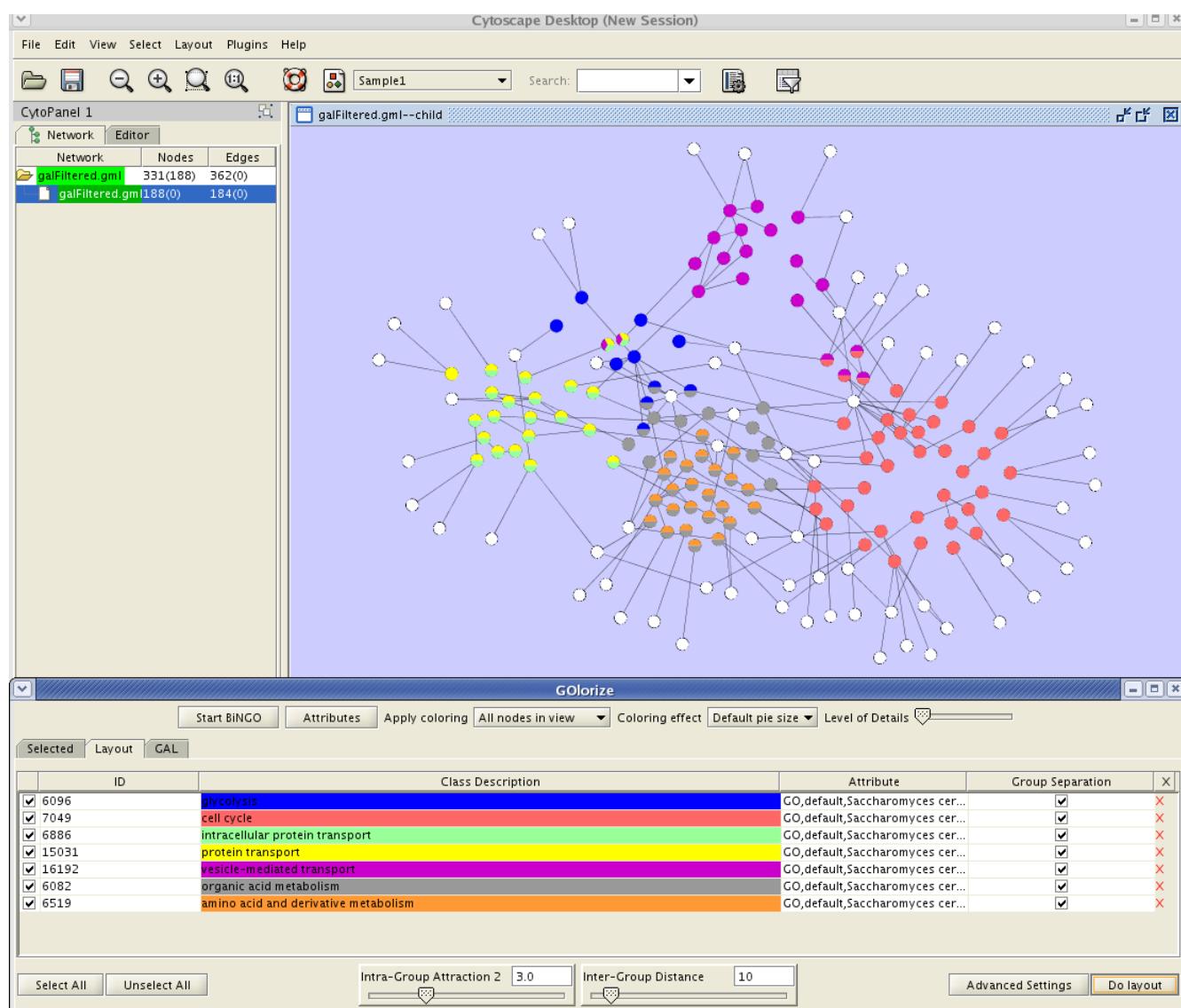
Our modified force-directed layout algorithm finds the placement of the nodes based on both their connection structure (the original edges) and class structure (the selected GO categories). Globally, the operation of the class-directed layout algorithm is organized through the following three phases:

1. Initial node placement using force-directed optimization, where in addition to the standard attractive forces between each connected node, an extra attractive force is applied to the nodes belonging to the same class. The extra attraction is directed by adding virtual edges between class members and a virtual class node representing the particular class. This phase finds good initial positions for the class nodes.
2. Subsequent separation of the classes by moving the class nodes in the same proportion away

from the center of gravity of all nodes. This phase aims at providing maximal distinction between the different classes, while still preserving the relative placement of the nodes obtained in the initial layout phase.

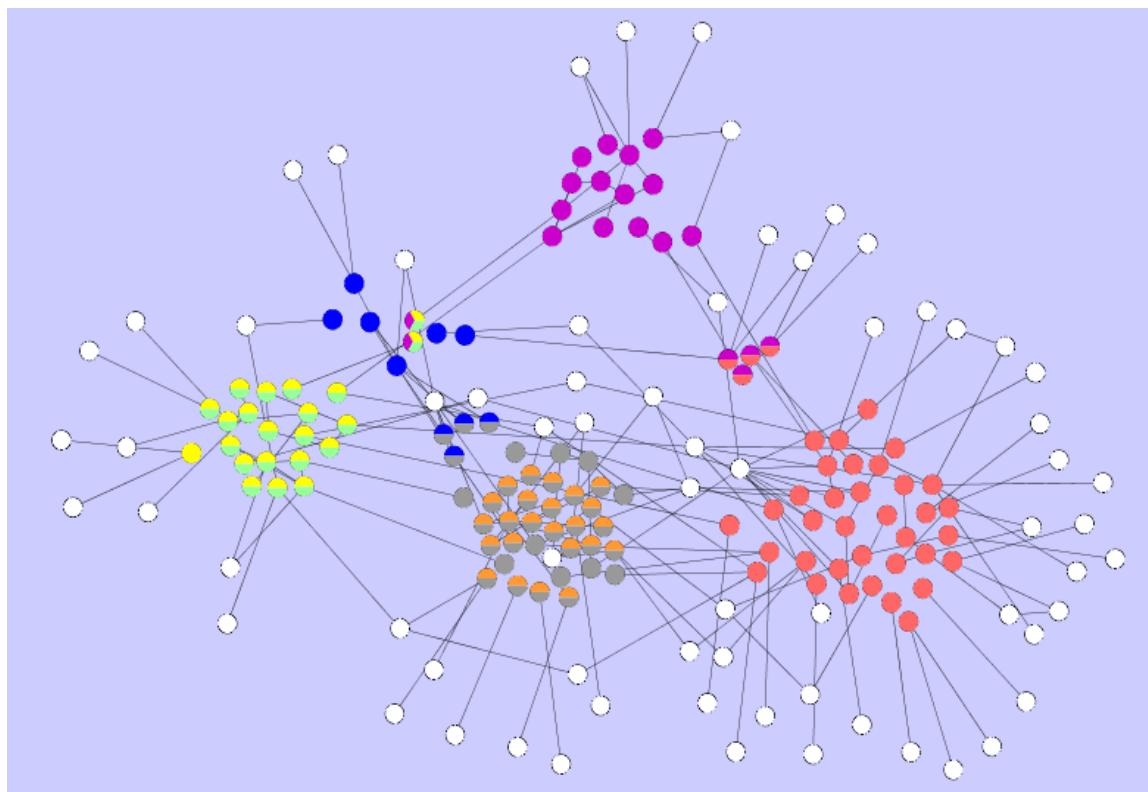
3. Final layout phase uses the same force-directed optimization process as in the step 1, but with class nodes fixed to the positions determined in the step 2. The aim of the step is to fine-tune the placement of the actual nodes. Neither the virtual edges nor the class nodes are shown in the final visualization.

In the Layout tab, the user can specify the parameters of the above layout algorithm. The two key parameters are the strength of the attraction within a class in the layout phase 3 (termed Intra-Group Attraction 2) and the extent of which the class nodes are moved in the separation step 2 (Inter-Group Distance). The example layout shown below corresponds to the default values of these two parameters (3 and 10), after pressing the Layout button. Again, the results can be different between two Cytoscape sessions, and even within the same session, due to random starting positions and movements.

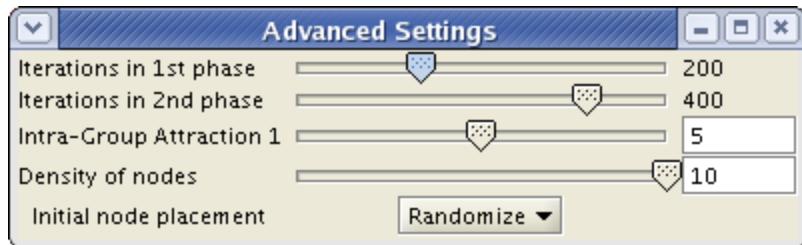


Nodes that belong to the same class (indicated by the same color) are grouped together, and the nodes with multiple GO class memberships (pie coloring) lie typically in between their main classes. Unclassified nodes (white color) are placed according to their connection structure only (the edges). The Layout tab also allows user to control whether the placement of a class node in the in the final layout phase 3 is free or fixed to the location determined in the separation step 2. This is specified using the check-boxes under the column Group Separation, and it can help to decide placements for small or heavily overlapping node classes.

A nice feature of the layout algorithm is that it is capable of grouping the class members close to each other even if the original network was disconnected, especially if the Intra-Group Attraction is substantially larger than one. This is because the parameter is directly proportional to the standard attraction between connected nodes. Increasing the value of Intra-Group Attraction, and decreasing the value of Inter-Group Distance respectively, results in more compact node classes (an example below). Such a layout emphasizes the connections between the classes (or metanodes), while the within-class connections are not so clearly visible anymore.



In the Advanced Settings, the user can adjust also several other parameters, including the number of iterations performed in the two layout phases, the strength of the attraction within a class in the initial layout phase 1, and the strength of the standard attraction between two connected nodes (Density of nodes). Two alternative modes to the initial placement of the nodes is implemented – In the first one, all the nodes start from random positions, whereas in the second one, the standard and class nodes are initially placed on two circles within each other (this provides reproducible solutions within a session).

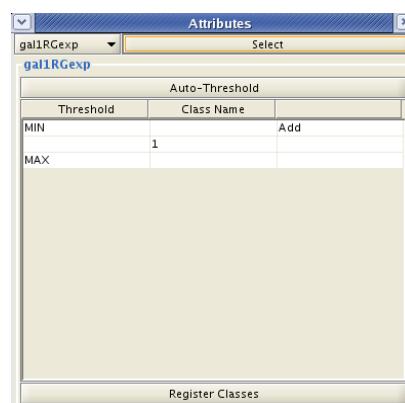


A stable solution can be obtained even for a large network consisting of hundreds of nodes in a few seconds on a standard workstation. The fast operation of the layout algorithm makes it possible to experiment with different parameter combinations and different starting positions. The aim of the tool is not to enable a fully automated visualization tool, but rather to facilitate in discovering whether or not there is an intrinsic GO class structure hidden in the network of original interactions. In some cases, such a structure can not be discovered even after trying a multitude of parameter combinations.

### Step 6 – Using other attributes

This version of GOLORIZE plug-in can deal with several types of attributes/values, in addition to ontologies/annotations, when defining the node classes for the visualization. Therefore, the annotation and term columns in the tables have been renamed as Attribute and Class description, respectively. Three types of node attributes can be used depending on the data type: continuous attributes handle integer or floating numbers such as expression data from e.g. microarray or mass-spectrometry experiments; string attributes can handle e.g. node labels such as lethal or non-lethal gene or protein; and list of strings attributes can be used as a generic way to define classes based on e.g. external files.

To test this feature, import an expression data by going to File->Import->Attribute/Expression Matrix and choose the file galExpData.pvals, which contains the expression changes together with their associated p-values in the particular experiment. Click on the Attributes button in the GOLORIZE panel. At the upper left corner of the new panel is a pull-down menu listing all the attributes loaded to Cytoscape. Choose the attribute gal1RGexp, which represents the expression changes in response to altered Gal1 transcription factor, and validate this choice by clicking on the Select button. The Attributes panel for expression data-based class definition should now look like the screen-shot below.



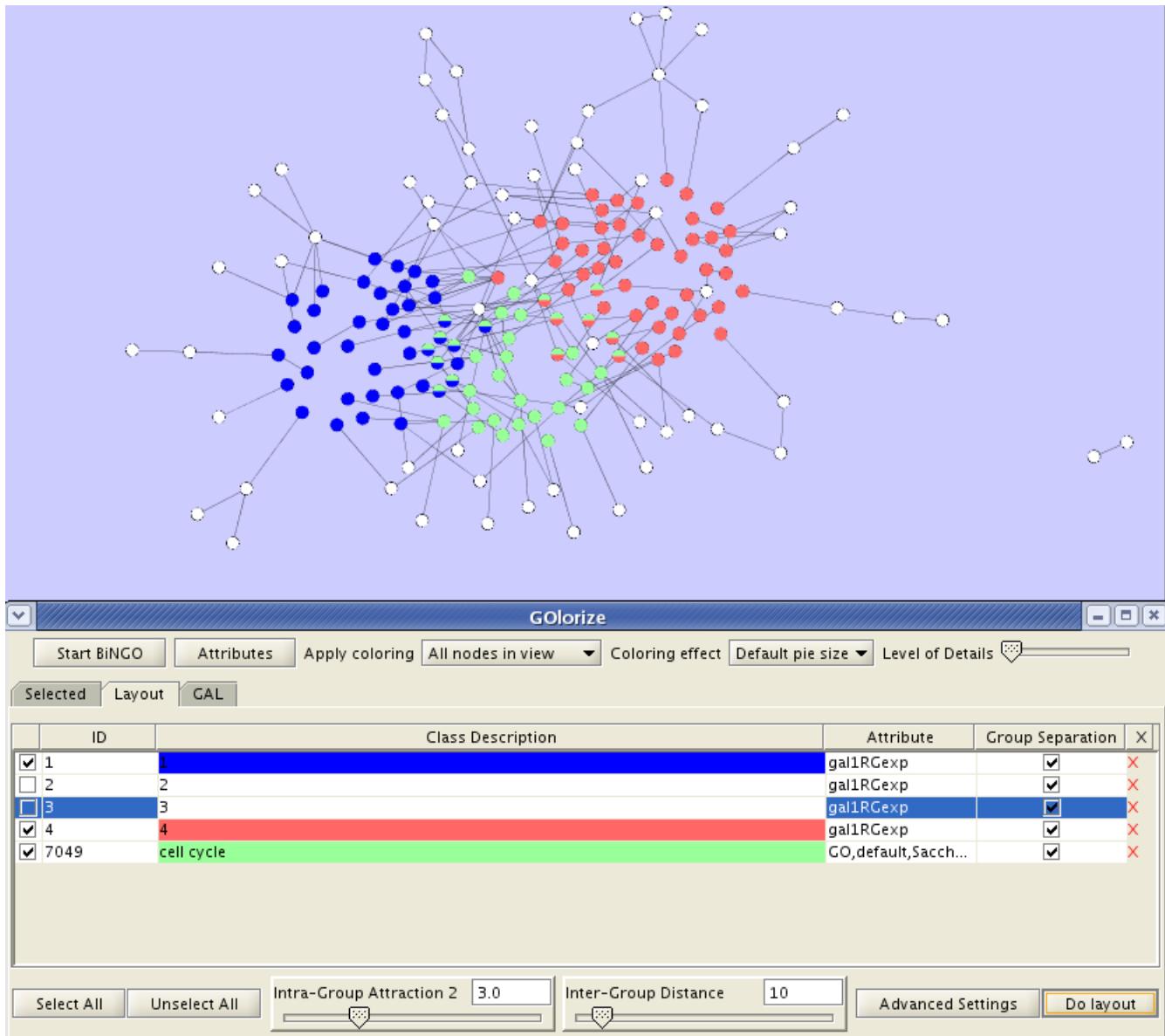
The class definition mechanism recognizes automatically that the selected node attribute, gal1RGexp, contains floating numbers. The Threshold column shows the values that separate two consecutive classes, and the Class Name column numbers the corresponding classes. The thresholds can be adjusted by double-clicking the cells and typing in new value. By clicking on the cells of the last column, the user can add or remove threshold values resulting in a split or merge of the corresponding classes. By default, only one class is defined that covers the whole range of the particular attribute (from minimum to maximum, see the example above).

The below example was created by clicking on the Add cell three times and pressing the Auto-Threshold button. These data-driven thresholds are based on equally distributed quantiles evaluated from the nodes present in the selected network, not from the total set of nodes that have a value associated with the gal1RGexp attribute (more than 6000 proteins). The class definitions can be validated by clicking on the Register Classes button. This adds four new classes into the GOlorize Selected tab. These classes can be utilized in the same way as the node classes obtained with BiNGO over-representation analysis.

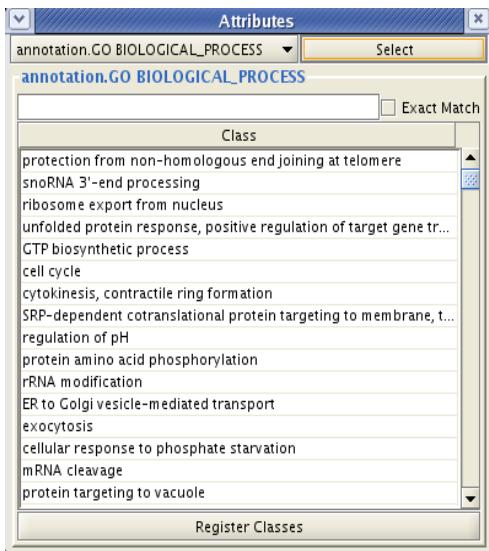
Auto-Threshold		
Threshold	Class Name	
MIN		Add
-0.209	1	Remove
-0.0115	2	Add
-0.0115	3	Remove
0.165	4	Add
MAX		

Register Classes

The below simple example demonstrates how the four classes defined on the basis of the expression data can be used together with GO categories (cell cycle in the sample case) in the network visualization. Color all the five classes with the Auto-Colors option, and layout the selected network without taking into account the less significant classes 2 and 3. Blue color identifies now under-expressed nodes, and red the over-expressed ones. The purpose of this ad-hoc example was just to show that different types of attributes and class definitions can be freely mixed in the GOlorize layout.



The final sample use case defines classes from string attributes. This example is done in the case of gene ontology annotations, which are managed in Cytoscape as lists of string attributes for nodes. However, any other string attribute could be used instead, e.g., imported attributes from an external file in which case the node classification can be freely defined by the user. The list of strings attributes makes it possible to define also overlapping node classes or clusters. Go to the import menu File->Import->Ontology and Annotation, select the Gene association file for *Saccharomyces cerevisiae* and Gene Ontology Full. After pressing the Import button, the new string attributes are available in the Golorize attributes window. Choose e.g. annotation.GO BIOLOGICAL\_PROCESS and press Select, and the following new table should appear that lists all the distinct strings for the attribute:



Type e.g. “mito” in the text field and press enter. The resulting table includes all the GO terms that contain the particular string. By entering several terms subsequently, the user can constraint the class definition further. Double clicking on one of the annotations defines the class based on the single string only (exact match option). Registration adds the new class into the GOLORIZE Selected tab. Note that the ontology structure is not taken into account when defining the class membership of a node based on its attributes. This means that even if an ancestor of a term match the query string, but not the term itself, the corresponding node does not belong to the particular class. As the BiNGO classes preserves the hierarchy of the terms, the two mechanisms for defining node classes can be considered as complementary to each other.